

# Recovering rank-one matrices via rank- $r$ matrices relaxation

Pengwen Chen<sup>\*</sup>, Hung Hung<sup>†</sup>

May 8, 2014

## Abstract

PhaseLift, proposed by E.J. Candès et al., is one convex relaxation approach for phase retrieval. The relaxation enlarges the solution set from rank one matrices to positive semidefinite matrices. In this paper, a relaxation is employed to nonconvex alternating minimization methods to recover the rank-one matrices. A generic measurement matrix can be standardized to a matrix consisting of orthonormal columns. To recover the rank-one matrix, the standardized frames are used to select the matrix with the maximal leading eigenvalue among the rank- $r$  matrices. Empirical studies are conducted to validate the effectiveness of this relaxation approach. In the case of Gaussian random matrices with a sufficient number of nearly orthogonal sensing vectors, we show that the singular vector corresponding to the least singular value is close to the unknown signal, and thus it can be a good initialization for the nonconvex minimization algorithm.

## 1 Introduction

Phase retrieval is one important inverse problem that arises in various fields, including electron microscopy, crystallography, astronomy, and optics. [17, 14, 16, 19, 10, 9]. Phase retrieval aims to recover signals from magnitude measurements only (optical devices do not allow direct recording of the phase of the electromagnetic field).

Let  $x_0 \in \mathbf{R}^n$  or  $x_0 \in \mathbf{C}^n$  be some nonzero unknown vector to be measured. Let  $A \in \mathbf{R}^{N \times n}$  be the matrix whose rows are sensing vectors  $\{a_i \in \mathbf{R}^n\}_{i=1}^N$  or  $\{a_i \in \mathbf{C}^n\}_{i=1}^N$ . The measurement vector  $b \in \mathbf{R}^N$  is the magnitude,

$$b = |Ax_0|, \text{ or } b_i = |a_i \cdot x_0| \text{ for } i = 1, \dots, N.$$

Obviously, the signal  $x_0$  can be determined up to a global phase factor at best, i.e., because

$$|x_0 \cdot a_j e^{i\theta}| = |x_0 \cdot a_j| \text{ for any } \theta \in [0, 2\pi],$$

<sup>\*</sup>Applied Mathematics, National Chung Hsing University, Taiwan

<sup>†</sup>Institute of Epidemiology and Preventive Medicine, National Taiwan University, Taiwan

then  $x_0 e^{i\theta}$  is also a solution. The recovery of  $x_0 e^{i\theta}$  is referred to as the exact recovery. When  $A$  is a Fourier matrix, the problem is known as phase retrieval. With this specific measurement matrix, the task becomes more demanding, because Fourier magnitude is not only preserved under global phase shift, but also under spatial shift and conjugate inversion, which yields twin images[9].

The first widely accepted phase retrieval algorithm was presented by Gerchberg and Saxton[12]. Fienup[11] developed the convergence analysis of the error-reduction algorithm and proposed input-output iterative algorithms. The basic and hybrid input-output algorithms can be viewed as a nonconvex Dykstra algorithm and a nonconvex Douglas-Rachford algorithm, respectively[3]. Empirically, the hybrid input-output algorithm is observed to converge to a global minimum (no theoretical proof is available)[19].

The major obstacle to phase retrieval is caused by the lack of convexity of the magnitude constraint[9]. PhaseLift[6], proposed by E.J. Candès et al., is one convex relaxation approach for phase retrieval. The relaxation changes the problem of vector recovery into a rank-one matrix recovery. The global optimal solution can be achieved, when  $A$  is a Gaussian random matrix and  $N \geq Cn$  with some absolutely constant  $C$ [5]. To some extent, this approach provides a solution to the phase retrieval problem, at least from the theoretical perspective, provided that the feasible set can shrink to one single point under a sufficient number of measurements. In practice, the sensing matrix  $A$  does not belong to this specific Gaussian model or uniform models, and the computational load of solving the convex feasibility problem can be too demanding. In particular, it requires the computation of all the singular values in each iteration.

In this paper, we explore the possibility of using the rank- $r$  matrix relaxation in phase retrieval. In the first section, to illustrate the idea, we review the exact recovery condition in PhaseLift. Typically, the exact recovery of rank-one matrices requires a large  $N/n$  ratio. We *standardize* the frame, such that each matrix in the feasible set has an equal trace norm. Then, the desired rank one matrix is the matrix whose leading eigenvalue is maximized. Gradually enlarging the leading eigenvalue, the matrix moves towards the rank one matrix with high probability. Our simulation result substantiates the effectiveness of recovering rank one matrices.

To reduce the computational load, in section 2, we apply the relaxation to the nonconvex alternating direction minimization method (ADM) proposed in [22]. Frames are standardized to ensure the equal trace among all feasible solutions. In theory, searching for the optimal solution in a higher dimensional space can alleviate the stagnation of local optima. Finally, with a sufficient amount of nearly orthogonal sensing vectors, we show that the corresponding singular vector is close to the unknown signal and can thus be a good initialization. To some extent, this theoretical result provides a partial answer to the solvability of phase retrieval. In fact, when there is a lack of nearly orthogonal sensing vectors, the ADM can fail to converge, as discussed in Section 3.1.

In section 3, we conduct a few experiments to demonstrate the performance of the ADM methods, including the convergence failure of noncon-

vex ADM, the comparison between rank one ADM to rank- $r$  ADM, and the application of phase retrieval computer simulations. Finally, given a generic matrix, we can find an equivalent matrix whose columns are orthogonal and whose rows have equal norm. We discuss the existence and uniqueness proof of the orthogonal factorization in the appendix.

## 1.1 Notation

In this paper, we use the following notations. Let  $x^\top$  be the Hermitian conjugate of  $x$ , where  $x$  can be real or complex matrices (or vectors). Hence,  $x$  is Hermitian if  $x = x^\top$ . The notation  $x^*$  is reserved for a limit point of a sequence  $\{x^k\}_{k=0}^\infty$  or the final iteration of  $x$  in the computation. Let  $\|x\|_F$  be the Frobenius norm. The function  $\text{diag}(X)$  produces a vector that is the diagonal of a matrix  $X$ . The pseudo-inverse of matrix  $X$  is denoted by  $X^\dagger$ . The vector  $e$  is a vector consisting of one, and  $e_j$  is the vector consisting of zero, except one at the  $j^{\text{th}}$  entry. Let  $x_0 \in \mathbf{R}^n$  be the unknown signal and  $A \in \mathbf{R}^{N \times n}$  or  $\in \mathbf{C}^{N \times n}$  be the sensing matrix. Hence  $N$  is the number of measurements.

## 1.2 Ratio $N/n$

We shall briefly outline the threshold ratio  $N/n$  on the exact recovery of  $x_0$  [1]. The result can be regarded as a worst-case bound, because we demand the exact recovery for all possible nonzero vectors  $x$ . Denote a nonlinear map associated with  $A$  by  $M^A : \mathbf{R}^n \rightarrow \mathbf{R}^N$ ,

$$M^A(x) = \sum_{k=1}^N |a_k \cdot x| e_k.$$

The range of the mapping  $M^A$  consists of all the possible measurement vectors  $b$  via the sensing matrix  $A$ .

Throughout this paper, we assume that  $A$  has rank  $n$ . We say that a matrix  $A \in \mathbf{R}^{N \times n}$  satisfies the *rank\* condition* if all square  $n$ -by- $n$  submatrices of  $A$  has full rank and  $N > n$ . That is, any  $n$  row vectors of  $A$  are linearly independent.

**Proposition 1.1.** *Suppose that  $A$  satisfies the rank\* condition. If  $N \geq 2n - 1$ , then  $M^A : \mathbf{R}^n \rightarrow \mathbf{R}^N$  is injective.*

*Proof.* Suppose that  $M^A(x) = M^A(\hat{x})$  with  $x \neq \hat{x}$ ; then  $|a_k \cdot x| = |a_k \cdot \hat{x}|$ . Rearrange the indices and assume

$$a_k \cdot x = a_k \cdot \hat{x} \text{ for } k = 1, \dots, l,$$

$$a_k \cdot x = -a_k \cdot \hat{x} \text{ for } k = l + 1, \dots, N.$$

Because  $N \geq 2n - 1$ , then either  $l \geq n$  or  $N - l \geq n$ . Suppose  $l \geq n$ . Then  $x - \hat{x} \in \mathbf{R}^n$  is orthogonal to  $a_1, \dots, a_l$ . The full rank condition yields  $x - \hat{x} = 0$ , which shows the nonexistence of two distinct vectors  $x, \hat{x}$ . Similar arguments apply to the case  $N - l \geq n$ .  $\square$

According to the above proof, when  $N \leq 2n - 2$ , we can find a pair of vectors  $x, \hat{x}$  such that  $|Ax| = b = |A\hat{x}|$ . Indeed, when  $N = 2n - 2$ , let  $u$  be the vector orthogonal to  $\{a_i\}_{i=1}^{n-1}$  and  $v$  be the vector orthogonal to  $\{a_i\}_{i=n}^{2n-2}$ . Then  $x = u + v$  and  $\hat{x} = u - v$  are the desired pair of vectors. However, for any particular vector  $x$ , it is possible that no  $\hat{x} \in \mathbf{R}^n$  exists in the case  $n + 1 \leq N \leq 2n - 2$ .

**Proposition 1.2.** *Fix  $x_0 \in \mathbf{R}^n$ . Suppose each row  $a_i$  of  $A$  is independently sampled from some continuous distribution on the unit sphere in  $\mathbf{R}^n$ . Let  $b = |Ax_0|$ . Then, with probability one,  $|Ax| = b$  has a unique solution  $x = x_0$  for  $N \geq n + 1$ .*

*Proof.* Assume  $N = n + 1$ . Write

$$A := \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \text{ with } A_1 \in \mathbf{R}^{n,n}, A_2 \in \mathbf{R}^{1,n}.$$

Then with probability one,  $A_1$  is full rank and thus we can find a unique nonzero vector  $c \in \mathbf{R}^n$  such that  $a_{n+1} = c^\top A_1$ . Clearly,  $c$  is a continuous random vector that depends on  $A_2$ .

Suppose that  $x = \hat{x}$  is another solution of  $|Ax| = b$ . Then  $\hat{x}$  should be one solution of  $2^n$  possible systems

$$\{(A_1 \hat{x})_i = \pm b_i\} \text{ for } i = 1, \dots, n.$$

Let  $y := A_1(x_0 \pm \hat{x})$ . Then,  $y$  must be one of the  $2^n$  vectors with  $y_i = \pm 2b_i$  or 0 for  $i = 1, \dots, n$ . Note that  $y$  is independent of the selection of  $A_2$ . Alternatively,  $a_{n+1}x_0 = \mp a_{n+1}\hat{x}$  yields the orthogonality between  $c$  and  $A_1(x_0 \pm \hat{x})$ , i.e.,

$$a_{n+1} \cdot (x_0 \pm \hat{x}) = c^\top A_1(x_0 \pm \hat{x}) = c \cdot y = 0.$$

Since  $c$  is a continuous random vector that depends on  $A_2$ , then with probability one,  $c \cdot y = 0$  leads to  $y = 0$ , which implies that  $x = \pm \hat{x}$  ( $A_1$  is full rank).  $\square$

However, for generic complex frames, the map is injective if  $N \geq 4n - 2$ , i.e., all vectors  $x_0 \in \mathbf{C}^n$  can be recovered. To recover a fixed vector  $x_0$ ,  $N \geq 2n$  is a necessary condition. Interested readers are referred to the discussion in [1] and [2].

One naive thought is that as  $N$  grows faster than the speed of  $n$ , the rank one matrix can be recovered. Unfortunately, this can be incorrect in some circumstances. We can construct some matrix  $A \in \mathbf{R}^{N \times n}$  with  $N$  being order of  $2^n$ , but some vector  $x_0$  still cannot be recovered due to the failure of the rank\* condition, see the following remark.

*Remark 1.3.* (Bernoulli random matrices) We construct an example, in which  $x_0 = e_1$  cannot be recovered from the measurement  $|Ax_0| = b = e$ . Denote by  $S \subset \mathbf{R}^n$  a set of vectors whose entries are  $\pm 1$ . There are  $2^n$  vectors in  $S$ . Pick any subset of  $N$  vectors from  $S$  as  $\{a_i\}_{i=1}^N$  ( $A \in \mathbf{R}^{N \times n}$  is the Bernoulli random matrix). All the vectors  $e_j, j = 1, \dots, n$  satisfy  $|Ae_j| = |Ax_0| = b = e$ . Since these matrices are indistinguishable,  $M^A$  does not have the injective property. Note that the rank of the random matrix  $A$  is  $n$  in most cases. The rank  $n$  condition on  $A$ , together with a large  $N$ , does not imply the exact recovery of  $x_0$ . One can easily verify that the Fourier matrix yields the same difficulty.

### 1.3 PhaseLift

Next, we introduce the PhaseLift method proposed by Candès et al.[6]. To simplify the discussion, we focus on the noiseless case. Introduce the linear operator on Hermitian matrices,

$$\mathcal{A} : \mathcal{H}^{n \times n} \rightarrow \mathcal{R}^N, \mathcal{A}(X) := \text{diag}(A^\top X A) = b^2, b_i^2 = a_i^\top X a_i, i = 1, \dots, N.$$

An equivalent condition of  $|Ax_0| = b$  is that  $X := x_0 x_0^\top$  is a rank-one solution to  $\mathcal{A}(X) = b^2$ . Hence, the phase retrieval problem can be formulated as the matrix recovery problem,

$$\min_X \text{rank}(X) \text{ subject to } \mathcal{A}(X) = b^2, X \succeq 0.$$

By factorizing a rank one solution of  $X$ , we can recover the signal  $x_0$ .

To overcome the difficulty of rank minimization, Candès et al. [6] propose a convex relaxation of the rank minimization problem, which is the trace minimization problem,

$$\min_X \text{tr}(X), \text{ subject to } \mathcal{A}(X) = b^2, X \succeq 0.$$

When

$$I_{n \times n} \in \text{span}\{a_i a_i^\top\}_{i=1}^n,$$

the condition  $\mathcal{A}(X)_i = \text{tr}(a_i a_i^\top X)$  automatically determines the trace  $\text{tr}(X)$  of  $X$  and then the trace minimization objective is redundant. Recovering  $X_0 = x_0 x_0^\top$  can be achieved via solving the following convex feasibility problem,

$$\{X : X \succeq 0, \mathcal{A}(X) = b^2\}. \quad (1.1)$$

In the next subsection, we will show that we can always remove the trace minimization objective via an orthogonal decomposition on  $A$ , either SVD or QR factorizations.

The following Prop. [6] illustrates the optimality of the feasibility problem, which is a key tool for justifying the exact recovery theoretically. The proof can be found in [8].

**Proposition 1.4.** *Suppose that the restriction of  $\mathcal{A}$  to the tangent space at  $X_0 := x_0 x_0^\top$  is injective. One sufficient condition for the exact recovery is the existence of  $y \in \mathbf{R}^N$ , such that*

$$Y := \mathcal{A}^\top y = A^\top \text{diag}(y) A = \sum_{i=1}^N y_i a_i a_i^\top$$

satisfies

$$Y_T = 0 \text{ and } Y_{T^\perp} \succ 0.$$

The proposition states one sufficient condition under which  $x_0 x_0^\top$  can be recovered from the frame  $A$ . In the real case, when  $N \geq 2n - 1$ , the rank\* condition on  $A$  is one sufficient condition to ensure the injective property of the restriction of  $\mathcal{A}$ . Indeed, for any  $x_0 \neq 0$ ,  $Ax_0$  consists of at most  $n - 1$  zeros thanks to the rank\* condition, thus  $Ax_0$  consists of at least  $n$  nonzero entries. Since the tangent space at  $x_0 x_0^\top$  consists of  $\hat{X}$  in a form  $x_0 x^\top + x x_0^\top$  with some  $x \in \mathbf{R}^n$ , then

$$\mathcal{A}(\hat{X}) = \mathcal{A}(x_0 x^\top + x x_0^\top) = 2(Ax_0)(Ax) = 0 \in \mathbf{R}^N$$

yields  $x = 0$  (due to the rank\* condition), which implies  $\hat{X} = 0$ .

## 1.4 Special frames

We shall highlight three special frames where the feasible set only consists of one single point. Thus the unknown signal  $x_0$  can be recovered via PhaseLift. In the first case, we show that a frame with  $N = n + 1$  measurement vectors are sufficient to determine the unknown matrix  $X = x_0 x_0^\top$ .

**Proposition 1.5.** *Suppose that  $a_j = e_j$  for  $j = 1, \dots, n$  and  $a_{n+1}(j)x_0(j) > 0$  for  $j = 1, \dots, n$ .<sup>1</sup> Then, the feasible set of PhaseLift consists of only one single point,  $x_0 x_0^\top$ .*

*Proof.* From  $a_j = e_j$ , we have

$$a_j^\top X a_j = X_{j,j} = |x_0(j)|^2.$$

The positive semidefinite requirement of  $X$  yields  $X_{i,j}^2 \leq X_{i,i} X_{j,j}$ . The measurement  $a_{n+1}^\top X a_{n+1} = |a_{n+1} \cdot x_0|^2$  enforces  $a_{n+1}^\top X a_{n+1}$  to reach its upper bound among  $X$  being positive semidefinite, i.e., the inequalities in the following relation become equalities,

$$\begin{aligned} a_{n+1}^\top X a_{n+1} &= \sum_{i,j} a_{n+1}(i) X_{i,j} a_{n+1}(j) \\ &\leq \sum_{i,j} |a_{n+1}(i) a_{n+1}(j)| \sqrt{X_{i,i}} \sqrt{X_{j,j}} \leq \left( \sum_j |a_{n+1}(j)| \sqrt{X_{j,j}} \right)^2 = |a_{n+1} \cdot x_0|^2, \end{aligned}$$

where we used the assumption  $a_{n+1}(j)x_0(j) > 0$  for all  $j = 1, \dots, n$ . Hence,  $X_{i,j}^2 = X_{i,i} X_{j,j}$  for all  $i, j$ , which implies  $X = x x^\top$  is the only feasible point with  $x := (a_{n+1} x_0 / |a_{n+1}|)$ .  $\square$

In the second case, the exact recovery is obtained via a set of sensing vectors orthogonal to  $x_0$ .

**Proposition 1.6.** *Suppose that some  $n - 1$  linear independent sensing vectors among  $\{a_i\}_{i=1}^n$  exist such that  $x_0 \cdot a_i = 0$ ; then, PhaseLift with measurement matrix  $A$  recovers the matrix  $x_0 x_0^\top$  exactly.*

*Proof.* Without loss of generality, assume that  $x_0 = e_1$  and write  $A$  as an  $n \times n$  matrix,

$$A = \begin{pmatrix} 1_{1 \times 1} & *_{1 \times (n-1)} \\ 0_{(n-1) \times 1} & A_1 \end{pmatrix}$$

where  $A_1 \in \mathbf{R}^{(n-1) \times (n-1)}$  has rank  $n - 1$ , i.e., it consists of linear independent columns. Choose  $y$  to be a vector with  $y_i > 0$  for  $i \geq 2$ . Then  $Y_T = 0$  and for any  $z \in \mathbf{R}^{n-1}$ ,

$$(A_1 z)^\top \text{diag}([y_2, \dots, y_n]) A_1 z = 0.$$

Because  $y_i > 0$  for all  $i = 2, \dots, n$ , then  $A_1 z = 0$ , i.e.,  $z = 0$ . Hence,  $Y_{T^\perp} \succ 0$ .  $\square$

---

<sup>1</sup> This condition states that the entries of  $a_{n+1}$  have the same sign as the ones of  $x_0$ .

This special choice of the first column of  $A$  indicates *the orthogonality between  $n - 1$  sensing vectors  $a_i$  and  $x_0$* . However, the orthogonality is generally not satisfied for arbitrary vector  $x_0$ .

In the third case, the exact recovery can be obtained via some structured sensing matrix, which in fact fails the rank\* condition ( $M^A$  is not injective).

**Proposition 1.7.** *Suppose that  $N = 2n - 1$  and the sensing vectors in  $A$  are  $a_i = e_i$  for  $i = 1, \dots, n$  and*

$$a_{n+i} = e_i + \beta_i e_{i+1} \text{ with } \beta_i \neq 0 \text{ for } i = 1, \dots, n - 1.$$

*Suppose that the entries of  $x_0$  are nonzero. Then PhaseLift with measurement matrix  $A$  recovers the matrix  $x_0 x_0^\top$  exactly.*

*Proof.* To simplify the discussion, assume  $x_0 = e$  and replace vectors  $a_i$  with vectors  $a_i x_0$  for all  $i$ . Any matrix  $X$  in the feasible set has the form,

$$X \in \mathbf{R}^{n,n} = \begin{pmatrix} 1 & 1 & & & \\ 1 & 1 & 1 & & \\ & 1 & 1 & 1 & \\ & & & \ddots & \\ & & & 1 & 1 \end{pmatrix},$$

i.e.,  $X_{i,i+1} = X_{i,i} = X_{i+1,i} = 1$ .

Claim: Because  $X$  is positive semidefinite, any principal sub-matrices of  $X$  are positive semidefinite, which implies that  $X = ee^\top$ .

Start with  $\alpha := X_{i_1, j_1} = X_{j_1, i_1}$  with  $i_1 = j_1 + 2$ . Consider the principal sub-matrix  $\{X_{i,j} : i, j \in \{j_1, j_1 + 1, j_1 + 2\}\}$ . Compute the determinant of this submatrix

$$-1 + 2\alpha - \alpha^2 = -(1 - \alpha)^2.$$

Hence, the nonnegative determinant yields  $\alpha = 1$ . Similar arguments work for  $i_1 = j_1 + 3, \dots$ . In the end, all entries of  $X$  must be 1, i.e.,  $X = x_0 x_0^\top$  is the only matrix in the feasible set.  $\square$

Readers can apply the similar arguments to the recovery of  $X_0 = x_0 x_0^\top$  with  $x_0 \in \mathbf{C}^n$  via the following matrix: Let  $A \in \mathbf{C}^{N,n}$  with  $N = n + 2(n - 1)$  and

$$a_i = e_i, \quad a_{n+i} = e_i + \beta_i e_{i+1}, \quad \text{and} \quad a_{2n-1+i} = e_i + \gamma_i e_{i+1}, \quad \text{where}$$

$$\beta_i \neq \gamma_i \text{ are nonzero for } i = 1, \dots, n - 1.$$

See [4] for more discussion on the usage of  $N = 3n - 2$  sensing vectors.

## 1.5 Reduction of $N/n$ via standardized frames

In the following, some orthogonality on  $A$ ,  $A^\top A = I_{n \times n}$  is expected to implement the matrix recovery algorithm. We say that a measurement matrix (a frame)  $A$  is *standardized* if  $A$  consists of orthonormal columns, i.e.,  $A^\top A = I_{n \times n}$ . In fact, given any measurement matrix  $A \in \mathbf{R}^{N \times n}$  with rank  $n$ , we can take the QR decomposition of the measurement matrix  $A$ ,  $A = QR$  with  $Q \in \mathbf{R}^{N \times n}$  consisting of orthonormal columns and

$R \in \mathbf{R}^{n \times n}$  being upper triangular. The rank of  $A$  is equal to the rank of  $R$ . Hence, denoting  $Rx$  by  $y$ , the problem is reduced to solving  $y$  from the measurements

$$|Ax| = |QRx| = |Qy| = b.$$

Once  $y$  is obtained,  $x$  can be computed via simply inverting the matrix  $R$ . Hence, the original frame  $A$  is *equivalent* to the standardized frame  $Q$  in the sense that the two transforms  $A, Q$  have the same range. That is,  $M^A$  is injective if and only if  $M^Q$  is injective.

In the section, we propose one modification on PhaseLift to recover the rank one matrix  $X_0$ . The idea is based on the following simple fact. *Among the feasible set  $\mathcal{A}(X) = b^2$  and  $X$  being positive semidefinite, to recover the rank one solution, we should choose the matrix  $X$  whose leading eigenvalue is maximized.*

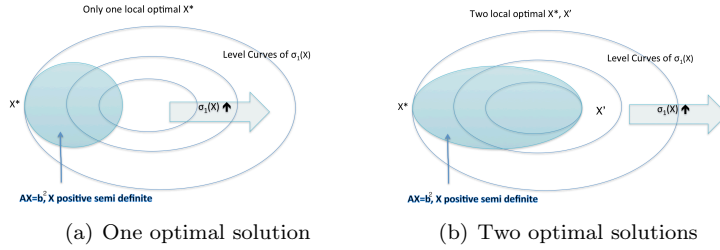


Figure 1: Maximizing the leading eigenvalue yields the rank one solution.

Consider the model

$$\min_X -\sigma_1(X),$$

subject to  $X$  positive semidefinite and  $\mathcal{A}X = b^2$ , where  $\sigma_1(X)$  refers to the largest eigenvalue function of  $X$ . See Fig. 1.

Then we have the following theoretical result.

**Theorem 1.8.** *Suppose that  $A$  is a standardized matrix with  $N \geq 2n - 1$  in the real case and  $N \geq 4n - 2$  in the complex case. Then with probability one, the global minimum occurs if and only if the minimizer  $X$  is exactly  $X = x_0 x_0^\top$ .*

*Proof.* Because  $|Ax|^2 = b^2$  and  $A$  consists of orthogonal columns,

$$e \cdot b^2 = \text{tr}(X) = \sum_{i=1}^n \sigma_i(X),$$

where  $\sigma_i(X)$  refers to the  $i$ -th eigenvalue of  $X$ . Because  $X$  is positive semidefinite, the largest eigenvalue of  $X$ , which cannot exceed  $e \cdot b^2$ , is maximized if and only if  $X$  is a rank one matrix. Finally, according to the above results, when  $N$  exceeds the thresholds  $2n - 1$  or  $4n - 2$ , with probability one, the rank one matrix is unique, which completes the proof.  $\square$



To address the problem, we propose the following alternating direction method (ADM). The ADM can be formulated as

$$\min_{X,Y} L_\beta(X, Y, \lambda) := -\sigma_1(X) - \lambda(X - Y) + \beta\|X - Y\|_F^2/2,$$

subject to  $X$  positive semidefinite and  $\mathcal{A}Y = b$ . Hence, the update of  $X, Y$  is

$$\begin{aligned} \arg \min_X -\sigma_1(X) + \beta\|X - Y - \lambda/\beta\|_F^2/2 \text{ subject to positive semidefinite,} \\ \arg \min_Y \beta\|X - Y - \lambda/\beta\|_F^2/2 \text{ subject to } \mathcal{A}Y = b^2. \end{aligned} \quad (1.2)$$

The iteration becomes

1. Update  $X$ : Write  $Y^k = UD_YU^\top$ ; then, thanks to the rotational invariance of Frobenius norm, the minimizer in Eq. (1.2) is

$$X^{k+1} = UD_XU^\top, \text{ where } D_X = \max(D_{Y+\lambda/\beta}, 0) + \beta^{-1}e_1e_1^\top,$$

where  $D_{Y+\lambda/\beta}$  is the diagonal matrix of the eigenvalue decomposition of  $Y^k + \lambda^k/\beta$  and the diagonal entries are in a decreasing order, i.e., the (1,1) entry is the largest eigenvalue and will be added by  $\beta^{-1}$  in the  $X$  update.

2. Update  $Y$  via the projection of  $X - \lambda/\beta$ ,

$$Y^{k+1} = \mathcal{A}^\top(\mathcal{A}\mathcal{A}^\top)^{-1}b + (I - \mathcal{A}^\top(\mathcal{A}\mathcal{A}^\top)^{-1}\mathcal{A})Z, \text{ where } Z = X^{k+1} - \lambda^k/\beta.$$

The matrix  $\mathcal{A}^\top(\mathcal{A}\mathcal{A}^\top)^{-1}\mathcal{A}$  is the orthogonal projector onto  $\text{Range}(\mathcal{A}^\top)$  which is spanned by  $\{a_i a_i^\top\}_{i=1}^N$ , each of which has trace one.

3. Update  $\lambda$ :

$$\lambda^{k+1} = \lambda^k - \beta(X^{k+1} - Y^{k+1}).$$

*Remark 1.9* (Counterexamples). Because  $\sigma_1(X)$  is convex in  $X$ , minimizing  $-\sigma_1(X)$  yields a non convex minimization problem. Theoretically, there is no guarantee that the global optimal solution can always be found numerically; however, the empirical study shows that the exact recovery will occur with high probability.

Here is one counterexample.

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \\ 1 & \sqrt{3}/2 & 0 \\ 1 & -\sqrt{3}/2 & 0 \\ 1 & 0 & \sqrt{3} \\ 1 & 0 & -\sqrt{3} \end{pmatrix}$$

Then the feasible set consists of matrices  $\begin{pmatrix} 1-3\mu & 0 & 0 \\ 0 & 2\mu & 0 \\ 0 & 0 & \mu \end{pmatrix}$  with  $\mu \in [0, 1/3]$ . The maximization of the leading eigenvalue leads to two possible solutions to  $|Ax|^2 = b^2$ : one is  $\mu = 0$  (the rank-one solution) and the other is  $\mu = 1/3$  (the rank-two solution), depending on the initialization. See Fig. 2.

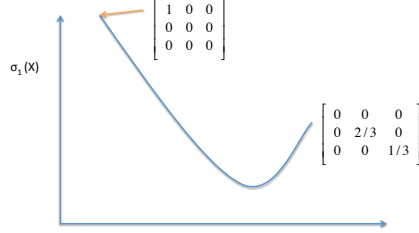


Figure 2: Maximizing the leading eigenvalue  $\sigma_1(X)$  yields two local optimal solutions.

The following experiments illustrate that when the ratio  $N/n$  is not large enough, the solution in PhaseLift is not rank one; we can successfully recover the rank one matrices via maximizing the leading eigenvalue; see Table 1 for the real case and Table 2 for the complex case.

Table 1: The number of successes out of 50 random trials with  $N = 2n - 1$ . “via Q” refers to the standardized measurement.

n	PhaseLift		$\min -\sigma_1(X)$	
	via A	via Q	via A	via Q
5	37	38	13	50
10	28	31	11	49
15	17	21	13	47
20	16	25	15	48
25	10	13	4	48
30	3	7	13	48
35	3	4	16	47
40	1	5	6	46
45	0	0	4	48
50	0	0	12	50

The result of nonconvex minimization depends on the choice of initialization for  $X^0$ . When  $X^0$  is near  $X_0$ , the exact recovery can be obtained.

**Proposition 1.10.** *Let  $X_0 = x_0 x_0^\top$ . Let  $X$  be some positive semidefinite matrix. Let  $f(t)$  be the spectral norm of matrices  $tX_0 + (1-t)X$ ,*

$$f(t) = \|tX_0 + (1-t)X\|, \quad 0 \leq t \leq 1.$$

*Let  $v_1$  be the unit eigenvector corresponding to the largest eigenvalue of  $X$ . If  $\text{tr}(X_0 v_1 v_1^\top) \geq \|X\|$ , then  $f(t)$  increases on the interval  $[0, 1]$ . Hence,  $\min(-\sigma_1(X))$  yields the recovery of  $X_0$ .*

*Proof.* Observe that  $f(t)$  is convex in  $t \in [0, 1]$ . It suffices to show that  $f(t) \geq f(0)$  for  $t \in (0, 1)$ . According to the subdifferential of the matrix

Table 2: The number of successes out of 20 random trials via  $N = 2n - 1, 3n - 1, 4n - 2$  standardized measurements ( complex case).

n	N=2n-1	N=3n-1	N=4n-2
5	3	20	20
10	2	20	20
15	1	20	20
20	0	20	20
25	0	20	20
30	0	20	20
35	0	20	20
40	0	20	20
45	0	20	20
50	0	20	20

spectral norm[21], we have

$$t^{-1}(f(t) - f(0)) \geq \text{tr}((X_0 - X)^\top G),$$

where  $G$  is a subgradient of the spectral norm at  $X$ . Choose  $G = v_1 v_1^\top$ , then we have  $\text{tr}((X_0 - X)^\top G) = \text{tr}(X_0 v_1 v_1^\top) - f(0)$ , which completes the proof.  $\square$

*Remark 1.11.* We provide a few examples to illustrate the recovery of  $X_0$  via  $-\min \sigma_1(X)$ . Let  $x_0 = e \in \mathbf{R}^3$  and  $X_0 = ee^\top \in \mathbf{R}^{3 \times 3}$ . Suppose  $a_1 = e_1, a_2 = e_2$  and  $a_3 = e_3$ . Then, any feasible matrix has the form

$$X = \begin{pmatrix} 1 & 1 - \alpha_1 & 1 - \alpha_2 \\ 1 - \alpha_1 & 1 & 1 - \alpha_3 \\ 1 - \alpha_2 & 1 - \alpha_3 & 1 \end{pmatrix} \text{ with } \alpha_i \geq 0.$$

Consider the case  $\alpha_1 = \alpha_2 = \alpha$  and  $\alpha_3 = \alpha t$ , then denote

$$X_{\alpha,t} := \begin{pmatrix} 1 & 1 - \alpha & 1 - \alpha \\ 1 - \alpha & 1 & 1 - \alpha t \\ 1 - \alpha & 1 - \alpha t & 1 \end{pmatrix}, \quad (1.3)$$

and

$$\det(X_{\alpha,t}) = \alpha^2 t(4 - 2\alpha - t).$$

Thus,  $X_{\alpha,t}$  is positive semidefinite if and only if

$$0 \leq t \leq 4 - 2\alpha \text{ and } 0 \leq \alpha \leq 2.$$

Hence,  $X_{\alpha,4-2\alpha}$  has two positive eigenvalues and one zero eigenvalue if  $0 < \alpha < 2$ . For instance, when  $\alpha = 1/2$  and  $t = 3$ ,  $X_{\alpha,t}$  has eigenvalues 1.5, 1.5, 0. From the previous proposition, a positive semidefinite matrix  $X_{\alpha,t}$  with  $0 \leq \alpha \leq 1/2$  can return to  $X_0$  via maximizing the leading eigenvalue. See Fig. 3.

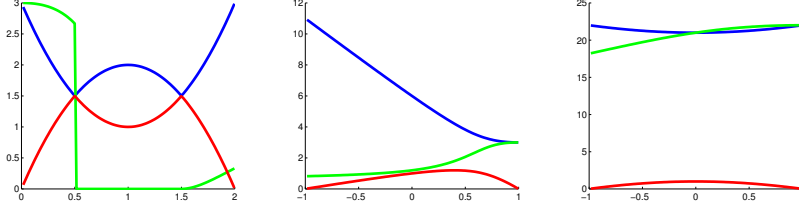


Figure 3: The left subfigure shows  $\sigma_1(X_{\alpha, 4-2\alpha})$  (blue),  $\sigma_2(X_{\alpha, 4-2\alpha})$  (red) and  $|x_0 \cdot v_1|^2$  (green). The middle and right subfigures show the results of  $X_{\alpha, 2\alpha-1}$  in Eq. (1.5) and  $X_{\alpha}$  in Eq. (1.7), respectively. The x-axis represents  $\alpha$  values.

The next example illustrates the necessity of trace invariance in recovering  $X_0$ . When the matrix trace is not constant in the feasible set, then maximizing the leading eigenvalue does not recover  $X_0$  in general. For instance, consider  $X_0 = x_0 x_0^\top$  with  $x_0 = e = [1, 1, 1]^\top$ , and

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 2 \end{pmatrix}, \quad b = |Ae| = \begin{pmatrix} 2 \\ 1 \\ 1 \\ 4 \end{pmatrix}.$$

Any matrix  $X$  in the feasible set has the form

$$X_{\alpha, \beta} := \begin{pmatrix} 3-2\beta & \beta & 2-\alpha \\ \beta & 1 & \alpha \\ 2-\alpha & \alpha & 1 \end{pmatrix} \quad \text{with } \det(X_{\alpha, \beta}) = -(2\alpha - \beta - 1)^2, \quad (1.4)$$

and thus  $\det(X_{\alpha, \beta}) \geq 0$  yields  $\beta = 2\alpha - 1$ . In fact, the feasible set consists of matrices

$$X_{\alpha, 2\alpha-1} = (1-t)\hat{e}\hat{e}^\top + tee^\top, \quad \hat{e} = [3, 1, 1]^\top, \quad t = (\alpha + 1)/2 \in [0, 1]. \quad (1.5)$$

Maximizing the leading eigenvalue yields the solution  $e\hat{e}$ , which is not  $X_0$ . Alternatively, consider the QR factorization,

$$A = QR = \begin{pmatrix} 1/\sqrt{2} & 0 & -1/\sqrt{3} \\ 0 & 1 & 0 \\ 0 & 0 & 1/\sqrt{3} \\ 1/\sqrt{2} & 0 & 1/\sqrt{3} \end{pmatrix} \begin{pmatrix} \sqrt{2} & \sqrt{2} & \sqrt{2} \\ 0 & 1 & 0 \\ 0 & 0 & \sqrt{3} \end{pmatrix},$$

which yields the problem instead,

$$b = |QRe| = |Qx_0|, \quad x_0 := Re = \begin{pmatrix} 3\sqrt{2} \\ 1 \\ \sqrt{3} \end{pmatrix}.$$

Then the feasible set  $\{X : \text{diag}(QXQ^\top) = b^2, X \succeq 0\}$  consists of

$$X_{\alpha} := \begin{pmatrix} 18 & 3\sqrt{2}\alpha & 3\sqrt{6} \\ 3\sqrt{2}\alpha & 1 & \sqrt{3}\alpha \\ 3\sqrt{6} & \sqrt{3}\alpha & 3 \end{pmatrix}, \quad \alpha \in \mathbf{R}. \quad (1.6)$$

Let  $x_0 = [3\sqrt{2}, -1, \sqrt{3}]^\top$  and  $\hat{x}_0 = [3\sqrt{2}, -1, \sqrt{3}]^\top$ . The feasible set consists of matrices

$$X_\alpha = (1-t)\hat{x}_0\hat{x}_0^\top + tx_0x_0^\top, \quad t = (\alpha+1)/2 \in [0, 1]. \quad (1.7)$$

When  $\alpha$  lies in  $(0, 1)$ , maximizing the leading eigenvalue of  $X_\alpha$  yields the exact recovery of  $X_0$ .

## 2 Low rank approaches

In PhaseLift, all eigenvalues of  $X$  may be computed in each iteration to be projected on the feasible set, consisting of positive semidefinite matrices with rank  $n$ . The projection obviously becomes a laborious task when  $n$  is large. Here we propose to replace the feasible set with a subset consisting of rank- $r$  matrices, where  $r$  is much smaller than  $n$ .

Write the positive semidefinite matrices  $X$  in PhaseLift as  $X = xx^\top$  with  $x \in \mathbf{R}^{n,r}$  or  $x \in \mathbf{C}^{n,r}$ . Then, the original constraint in PhaseLift becomes

$$b^2 = \mathcal{A}(X) = \text{diag}(AXA^\top) = |Ax|^2.$$

### 2.1 ADM with $r = 1$

Here we focus on the case  $r = 1$ . In section 2.11, we will discuss the case  $r > 1$ . When  $r = 1$ , we arrive at the problem,

$$\text{finding } x \in \mathbf{R}^n \text{ or } \mathbf{C}^n \text{ satisfying } |Ax| = b.$$

In [22], the framework

$$\min \frac{1}{2} \| |z| - b \|^2, \quad \text{subject to } z = Ax \quad (2.1)$$

is proposed to address phase retrieval. They introduce the augmented Lagrangian function

$$L(z, x, \lambda) = \frac{1}{2} \| |z| - b \|^2 + \lambda \cdot (Ax - z) + \frac{\beta}{2} \| Ax - z \|^2. \quad (2.2)$$

The algorithm consists of updating  $z$ ,  $x$ , and  $\lambda$  as follows.

**Algorithm 2.1.** *Initialize  $x^0$  randomly and  $\hat{\lambda}^0 = 0$ . Then repeat the steps for  $k = 0, 1, 2, \dots$*

$$\begin{aligned} z^{k+1} &= \frac{u}{|u|} \frac{b + \beta|u|}{1 + \beta}, \quad u = Ax^k + \beta^{-1}\lambda^k, \\ x^{k+1} &= A^\dagger(z^{k+1} - \beta^{-1}\lambda^k), \\ \lambda^{k+1} &= \lambda^k + \beta(Ax^{k+1} - z^{k+1}). \end{aligned}$$

Let us simplify the algorithm. Let  $P = AA^\dagger = QQ^\top$ . Assume that  $A$  has rank  $n$ . By eliminating  $x^k$ , the  $\lambda$ -iteration becomes

$$\beta^{-1}\lambda^{k+1} = (I - P)(\beta^{-1}\lambda^k - z^{k+1}).$$

Thus,  $A^\dagger\lambda^{k+1} = 0$ . In the end, we have the following algorithm.

**Algorithm 2.2.** Denote  $\hat{\lambda}^k = \beta^{-1}\lambda^k$ . Initialize  $x^0$  randomly and  $\hat{\lambda}^0 = 0$ . Compute  $z^0$ . Then, repeat the steps for  $k = 0, 1, 2, \dots$ ,

$$\begin{aligned} z^{k+1} &= \frac{u}{|u|} \frac{b + \beta|u|}{1 + \beta}, \quad u = Pz^k + \hat{\lambda}^k, \\ \hat{\lambda}^{k+1} &= (I - P)(\hat{\lambda}^k - z^{k+1}). \end{aligned}$$

*Remark 2.3* (Equivalence under right matrix multiplication). Note that the iteration is updated via  $P = QQ^\top$ , instead of  $A$ . The matrix  $R$  does not appear in the  $z$  and  $\lambda$  iterations. Thus, the algorithm is “invariant” with respect to  $R$ . That is, for any invertible matrix  $\hat{R}$ , we get the same iterations  $\{z^k, \hat{\lambda}^k\}_{k=1}^\infty$ , when  $(A, x^0)$  is replaced by

$$(Q\hat{R}, (\hat{R})^{-1}Rx^0).$$

In particular, the iteration with  $Q$  yields the same result as the one with  $A$  itself. However, ADM can produce different results when the *left* matrix multiplication on  $A$  is considered. See section 3.1.

Suppose that  $z^k$  converges to  $z^*$  and  $\hat{\lambda}^k$  converges to  $\hat{\lambda}^*$ . Then,  $Pz^* = z^*$  and  $P\hat{\lambda}^* = 0$ . Hence,  $x^* = A^\dagger z^*$ , and  $z^* = Pz^* = Ax^*$ . Consider the limit of the  $z$ -iteration,

$$(1 + \beta)z^* = \frac{u}{|u|}(b + \beta|u|) = \frac{u}{|u|}b + \beta(z^* + \hat{\lambda}^*).$$

Thus, we have the orthogonal projection of  $bu/|u|$  onto the range of  $A$  and its null space,

$$\frac{u}{|u|}b = z^* - \beta\hat{\lambda}^*,$$

$$\text{thus } \|b\|^2 = \|z^*\|^2 + \beta^2\|\hat{\lambda}^*\|^2. \quad (2.3)$$

This result shows  $\|Ax^*\| \leq \|b\|$ . In particular, when  $A = Q$ , we have  $\|x^*\| \leq \|b\| = \|x_0\|$ , i.e., any non-global solution has the smaller norm. Besides, Eq. (2.3) suggests the usage of smaller  $\beta$  to improve the recovery of  $x_0$ . Empirical experiments show that, starting with  $\lambda^0 = 0$ , a smaller value  $\beta$  leads to a higher chance of exact recovery.

To analyze the convergence, we write the function  $L(z, x, \lambda)$  as

$$\hat{L}(z, x, \lambda, s) := \frac{1}{2}\|z - bs\|^2 + \lambda \cdot (Ax - z) + \frac{\beta}{2}\|Ax - z\|^2,$$

where the entries of  $s$  satisfies  $|s_i| = 1$  for  $i = 1, \dots, N$  and clearly the optimal vector  $s$  to minimize  $\hat{L}$  is given by  $u/|u|$ . When  $s$  is fixed, then the following customized proximal point algorithm which consists of iterations

$$\begin{aligned} z^{k+1} &= s \frac{b + \beta|u|}{1 + \beta}, \quad u = Ax^k + \beta^{-1}\lambda^k, \\ \lambda^{k+1} &= \lambda^k + \beta(Ax^k - z^{k+1}), \\ x^{k+1} &= A^\dagger(z^{k+1} - \beta^{-1}\lambda^{k+1}) \end{aligned}$$

can be used to solve the least squares problem

$$\min \frac{1}{2}\|z - sb\|^2, \quad \text{subject to } z = Ax. \quad (2.4)$$

Gu et al. [13] provide the convergence analysis of the customized proximal point algorithm. More precisely, fixing  $s$ , let  $(z^*, x^*, \lambda^*)$  be a saddle point of  $\hat{L}(z, x, \lambda, s)$  and let

$$\|v^{k+1} - v^k\|_M^2 := (v^{k+1} - v^k)^\top M (v^{k+1} - v^k) \text{ with}$$

$$M := [\beta^{1/2}A, -\beta^{-1/2}I]^\top [\beta^{1/2}A, -\beta^{-1/2}I], \quad v := \begin{pmatrix} x \\ \lambda \end{pmatrix}.$$

In Lemma 4.2, Theorem 4.2 and Remark 7.1 [13], the sequence  $\{v^k\}$  satisfies

$$\|v^{k+1} - v^*\|_M^2 + \|v^k - v^{k+1}\|_M^2 \leq \|v^k - v^*\|_M^2,$$

and then  $\lim_{k \rightarrow \infty} \|v^k - v^{k+1}\|_M^2 = 0$ . Any limit point of  $[z^k, x^k, \lambda^k]$  is a solution of the problem in Eq. (2.4) with  $s$  fixed. However, the convergence analysis of the algorithm in Eq. (2.1) does not exist due to the lack of convexity in  $z$ . In fact, when  $s$  is updated in each  $z$ -iteration, this algorithm sometimes fails to converge, which is shown in our simulations; see Section 3.1.

## 2.2 Recoverability

We make the following two observations regarding  $|Ax_0| = b$ . Suppose the unknown signal  $x_0$  satisfies  $\|x_0\| = 1$ . First, the vector  $x$  is updated to maximize the inner product  $|Ax| \cdot b$  in the ADM. However, because the norm constraint  $\|x\| = 1$  is not enforced explicitly, a non-global maximizer  $x$  generally does not have the unit norm,  $\|x\| < \|x_0\| = 1$ . In fact, classic phase retrieval algorithms e.g., ER, BIO, HIO[11], do not enforce the constraint directly. Second, for those indices  $i$  with  $a_i \cdot x$  close to zero, the unit vector  $x$  to be recovered should be approximately perpendicular to these sensing vectors  $a_i$ . The candidate set  $\{x : |a_i \cdot x| \leq b_i\}$  forms a cone, including unit vectors approximately orthogonal to  $a_i$  corresponding to  $b_i$  close to zero. In particular, when  $a_1 \cdot x_0 \neq 0$  and  $a_i \cdot x_0 = 0$  for  $i = 2, \dots, N$  with  $N > n$ , then the cone is exactly the one-dimensional subspace spanned by the vector  $x_0$ .

One important issue of non-convex minimization problems is that the initialization can affect the performance dramatically. The  $x$ -iteration in the ADM tends to produce a vector close to the singular vector corresponding to its least singular value of  $A$ , in the sense that  $A^\dagger z$  boosts the component along the singular vector corresponding to the largest singular value of  $A^\dagger$ , i.e., the smallest singular value of  $A$ . In the following, we will analyze the recovery problem from a viewpoint of singular vectors and derive an error estimate between the unknown signal and the singular vector.

Rearrange the indices such that  $\{b_i\}$  are sorted in an increasing order,

$$0 \leq b_1 \leq b_2 \leq \dots \leq b_N.$$

Divide the indices into three groups,

$$\{1, 2, \dots, N\} = I \cup II \cup III.$$

We shall use subscripts  $I, II, III$  to indicate the indices from these three groups. The set  $I$  consists of the indices corresponding to the smallest  $N_I$  terms among  $\{b_i\}$ . The set  $II$  consists of the indices corresponding to the largest  $N_{II}$  terms among  $\{b_i\}$ . Denote the matrix consisting of rows  $\{a_i\}_{i \in I}$  by  $A_I$ . Let  $A_I$  and  $A_{II}$  consist of  $N_I$  and  $N_{II}$  rows, respectively. In the following, we illustrate that the singular vector  $x_{min}$  corresponding to the least singular value of  $A_I$  is a good initialization  $x^0$  in the ADM.

Without loss of generality, assume that  $x_0 = e_1$  and that all the rows  $\{a_i\}_{i=1}^N$  of  $A$  are normalized,  $\|a_i\| = 1$ . Observe that the desired vector satisfies  $|Ax_0| \leq b$  and  $\|x_0\| = 1$ . Hence, we look for a *unit* vector  $x$  in the closed convex set

$$|a_i \cdot x| \leq b_i \text{ for all } i.$$

Whether phase retrieval can be solved depends on the structure of  $A$ . We make the following assumptions.

- First, sufficiently many indices  $i \in I$  exist, such that

$$\|b_I\|^2 := \sum_{i \in I} b_i^2 \text{ is sufficiently small compared to } \|A_I x_1\|^2,$$

where  $x_1$  is a unit vector orthogonal to  $x_0$ . (Clearly, the matrix  $A_I$  has rank at least  $n - 1$ .)

- Second, there are at least  $n$  indices in  $II$ , such that

$$\text{entries } \{b_i\} \text{ are } \textit{large} \text{ for } i \in II,$$

and the matrix  $A_{II}$  has rank  $n$ .

The assumption that  $\{b_i\}_{i \in I}$  is close to zero implies that  $\{a_i\}_{i \in I}$  are almost orthogonal to  $x_0$ . Thus, we instead solve the problem

$$\min_x \|A_I x\|^2 \text{ with } \|x\| = 1.$$

The minimizer denoted by  $x_{min}$  is the singular vector  $x_{min}$  corresponding to the least singular value of  $A_I$ . Then,

$$\|A_I x_{min}\| \leq \|b_I\|.$$

Let  $\{0 \leq \mu_1 \leq \dots \leq \mu_n\}_{i=1}^n$  be the singular values of  $A_I$  with right singular vectors  $v_i$ . Then  $v_1 = \pm x_{min}$  and we can write

$$x_0 = \alpha_1 x_{min} + \sqrt{1 - \alpha_1^2} w,$$

with some unit vector  $w$  orthogonal to  $x_{min}$ . Let

$$x_1 := -(1 - \alpha_1^2)^{1/2} x_{min} + \alpha_1 w,$$

then  $x_1$  is a unit vector orthogonal to  $x_0$ . Note that

$$\|x_0 x_0^\top - x_{min} x_{min}^\top\|^2 = \|x_0 x_0^\top - x_{min} x_{min}^\top\|_F^2 / 2 = 1 - \alpha_1^2.$$

The following proposition gives a bound for the distance  $\|x_0 x_0^\top - x_{min} x_{min}^\top\|$  via the ratio  $\|A_I x_0\| / \|A_I x_1\|$ .



**Proposition 2.4.** *Let  $\alpha_1 = |x_0 \cdot x_{min}|$ . Then,*

$$(2 - \alpha_1^2)\|b_I\|^2 \geq (1 - \alpha_1^2)\|A_I x_1\|^2 \quad (2.5)$$

*Therefore, as  $\|b_I\|$  is small enough,  $1 - \alpha_1^2$  must be close to 0. Note that  $\|b_I\|^2 = \|A_I x_0\|^2$ , then*

$$\|b_I\|^2 \|A_I x_1\|^{-2} \geq \left( \frac{1 - \alpha_1^2}{2 - \alpha_1^2} \right) \geq \frac{1}{2} \|x_0 x_0^\top - x_{min} x_{min}^\top\|^2. \quad (2.6)$$

*Proof.* Note that

$$\begin{cases} \|A_I x_0\|^2 = \alpha_1^2 \|A_I x_{min}\|^2 + (1 - \alpha_1^2) \|A_I w\|^2, \\ \|A_I x_1\|^2 = (1 - \alpha_1^2) \|A_I x_{min}\|^2 + \alpha_1^2 \|A_I w\|^2. \end{cases}$$

Because  $x_{min}$  is the singular vector of the the least eigenvalue,

$$\begin{aligned} & (1 - 2\alpha_1^2) \|A_I x_0\|^2 - (1 - \alpha_1^2) \|A_I x_1\|^2 \\ &= \|A_I x_{min}\|^2 + 2(1 - \alpha_1^2)^2 (\|A_I w\|^2 - \|A_I x_{min}\|^2) \geq 0. \end{aligned}$$

□

Denote the sign vector of  $A_{II} x_{min}$  by  $u_{II}$ ,

$$u_{II} = \frac{A_{II} x_{min}}{|A_{II} x_{min}|}.$$

The closeness  $\|A_{II}(x_{min} - x_0)\|$  yields that  $A_{II} x_{min}$  should be close to  $A_{II} x_0$ . In particular, when the magnitude of entries  $b_{II}$  are large enough, both vectors have the same sign

$$\frac{A_{II} x_{min}}{|A_{II} x_{min}|} = \frac{A_{II} x_0}{|A_{II} x_0|}$$

and then

$$A_{II} x_0 = u_{II} b_{II}.$$

Once the sign vector is retrieved, the vector  $x_0$  can be computed via

$$x_0 = A_{II}^{-1}(u_{II} b_{II}).$$

## 2.3 Real Gaussian matrices

As an example of computing  $\|b_I\|$  and  $\|A_I x_1\|$ , let  $A \in \mathbf{R}^{N \times n}$  be a random matrix consisting of i.i.d. normal  $(0, 1)$  entries. In the following, we would illustrate that when  $N_I/N$  is small enough,  $x_{min}$  is a good initialization for  $x_0$  and thus we can recover the missing sign vector  $(Ax_0)/b$ .

Let  $x_0 = e_1$ . Then  $x := a_i \cdot x_0$  follows the distribution the normal  $(0, 1)$  distribution. Let  $a > 0$  be a function of  $N_I/N$  and satisfy

$$F(a) := \int_{-a}^a (2\pi)^{-1/2} \exp(-x^2/2) dx = N_I/N. \quad (2.7)$$

Then the leading terms of Taylor series of Eq. (2.7) yields

$$(2/\pi)^{1/2}(a - a^3/6) \leq N_I/N \leq (2/\pi)^{1/2}a. \quad (2.8)$$

Define the “truncated” second moment

$$\sigma_a^2 := \int_{-a}^a x^2 (2\pi)^{-1/2} \exp(-x^2/2) dx.$$

Taking the Taylor expansion of  $\sigma_a^2$  in terms of  $a$  yields the following result.

**Proposition 2.5.**

$$\sigma_a^2 \leq (2/\pi)^{1/2} a^3/3.$$

Additionally, when  $N_I/N$  tends to zero,  $\sigma_a^2$  is approximately  $(N_I/N)^3(\pi/6)$ , where Eq. (2.8) is used.

When  $N_I/N$  tends to zero, the following proposition shows that  $N_I^{-1/2}\|b_I\|$  is approximately  $(\pi/6)^{1/2}N_I/N$ , see Fig. 4 for one numerical simulation.

**Proposition 2.6.** Suppose  $N_I/N < c$  for some constant  $c$ . Then with high probability<sup>2</sup>

$$N_I^{-1/2}\|b_I\| \leq c_6(\pi/6)^{1/2}N_I/N,$$

for some constant  $c_6$ .

*Proof.* Because  $x_0 = e_1$ ,  $\|b_I\|$  is the norm of the first column of  $A_I$ . Let  $\hat{\mu} > 0$  be the sample  $N_I$  quantile [7],  $\hat{\mu} = |a_{N_I} \cdot x_0|$ . We have the following probability inequality for  $|\hat{\mu} - a|$ : For every  $\epsilon > 0$ ,

$$\mathbb{P}(|\hat{\mu} - a| > \epsilon) \leq 2 \exp(-2N\delta_\epsilon^2) \text{ for all } N,$$

where

$$\delta_\epsilon := \min\{F(a + \epsilon) - N_I/N, N_I/N - F(a - \epsilon)\}.$$

That is, with high probability we have

$$a - \epsilon \leq |a_{N_I} \cdot x_0| \leq a + \epsilon. \quad (2.9)$$

The proof is based on the Hoeffding inequality in large deviation; see Theorem 7, p. 10, [7]. Let  $\hat{N}_I$  be the cardinality of the set  $\hat{I} := \{i : |a_i \cdot x_0| \leq a\}$ . With high probability, we have<sup>3</sup>

$$\hat{N}_I \geq N_I(1 - \epsilon_1) \text{ for any } \epsilon_1 > 0.$$

Let  $\{Z_i\}_{i=1}^N$  be independent bounded random variables,

$$Z_i := (N/N_I)^3 ((a_i \cdot x_0)^2 - \sigma_a^2(N/N_I)) \text{ if } |a_i \cdot x_0| \leq a, \text{ zero, otherwise.}$$

$$\text{Then } \mathbb{E}[Z_i] = (N/N_I)^3 (\sigma_a^2 - \sigma_a^2(N/N_I)\mathbb{P}(|a_i \cdot x_0| \leq a)) = 0,$$

---

<sup>2</sup> An event occurs “with high probability” if for any  $\alpha \geq 1$ , the probability is at least  $1 - c_\alpha N^{-\alpha}$ , where  $c_\alpha$  depends only on  $\alpha$ .

<sup>3</sup> For every  $\epsilon_1 > 0$ , let  $\beta := F^{-1}(N_I(1 - \epsilon_1)/N)$  and  $\epsilon := a - \beta = F(N_I/N) - \beta > 0$ . Then

$$\beta - \epsilon \leq |a_{N_I/(1-\epsilon_1)} \cdot x_0| \leq \beta + \epsilon \text{ with high probability.}$$

i.e.,  $Z_i$  is centered. The Hoeffding inequality (e.g., Prop. 5.10 [20]) yields for some positive constants  $c_4, c_5$ ,<sup>4</sup>

$$\mathbb{P}(N^{-1} \left| \sum_{i=1}^N Z_i \right| \leq t) \geq c_4 \exp(-c_5 N t^2).$$

That is, with probability at least  $1 - c_4 \exp(-c_5 N t^2)$ ,

$$\left| N^{-1} \|\hat{b}_I\|^2 - \sigma_a^2 \frac{\hat{N}_I}{N_I} \right| \geq t(N_I/N)^3.$$

Thanks to  $\hat{N}_I \geq N_I(1 - \epsilon_1)$  with high probability and Eq. (2.8), (2.9),

$$\frac{\|b_I\|^2}{N_I} \leq \frac{\|\hat{b}_I\|^2}{\hat{N}_I} + \frac{\epsilon_1 N_I (a + \epsilon)^2}{\hat{N}_I} \leq \frac{\|\hat{b}_I\|^2}{\hat{N}_I} + \epsilon_1 \frac{(a + \epsilon)^2}{1 - \epsilon_1},$$

and

$$\begin{aligned} \frac{\|\hat{b}_I\|^2}{\hat{N}_I} &\leq \sigma_a^2 \frac{N}{\hat{N}_I} + t \frac{N_I}{\hat{N}_I} \left( \frac{N_I}{N} \right)^2 \\ &\leq (\sigma_a^2 (N_I/N)^{-3} + t/(1 - \epsilon_1)) \left( \frac{N_I}{N} \right)^2. \end{aligned}$$

Together with Prop. 2.5, with high probability

$$N_I^{-1/2} \|b_I\| \leq c_6 (\pi/6)^{1/2} (N_I/N) \text{ for some constant } c_6.$$

□

To bound the norm  $\|x_0 x_0^\top - x_{\min} x_{\min}^\top\|$  in Eq. (2.6), we need to compute  $\|A_I x_1\|^2$ . Denote by  $A'_I$  the sub-matrix of  $A_I$  with the first column deleted, i.e.,

$$A = [*_{N_I \times 1}, A'_I].$$

Denote by  $\{\delta_i\}_{i=2}^n$  the singular values of the sub-matrix  $A'_I$ . Since  $x_1$  is orthogonal to  $x_0$ , then we have lower bounds for  $\|A_I x_1\| = \|A'_I x_1\|$ , i.e.,  $\delta_2$ . Observe that the first column of  $A$  is independent of the remaining columns of  $A$ . Note that entries of  $A'_I$  are i.i.d. Normal(0,1). According to the random matrix theory of Wishart matrices, with high probability, the singular values  $\{\delta_i\}$  of the sub-matrix are bounded between  $\sqrt{N_I} - \sqrt{n}$  and  $\sqrt{N_I} + \sqrt{n}$ . More precisely,

$$\mathbb{P}(\sqrt{N_I} - \sqrt{n} - t \leq \delta_2 \leq \delta_n \leq \sqrt{N_I} + \sqrt{n} + t) \geq 1 - 2e^{-t^2/2}, \quad t \geq 0,$$

see Eq. (2.3) [18]. Together with  $N_I^{-1/2} \|b_I\| \leq c_6 (\pi/6)^{1/2} N_I/N$  in Prop. 2.6 and Eq. (2.6), we have

$$\begin{aligned} \|x_0 x_0^\top - x_{\min} x_{\min}^\top\|/\sqrt{2} &\leq \|b_I\|/(\sqrt{N_I} - \sqrt{n} - t) \\ &\leq c_6 \sqrt{\pi/6} (N_I/N) (1 - \sqrt{(n-1)/N_I} - \sqrt{t/N_I})^{-1}. \end{aligned}$$

Let  $c_7 := c_6 (1 - \sqrt{(n-1)/N_I} - \sqrt{t/N_I})^{-1}$ , then we have the following result.

---

<sup>4</sup> The sub-gaussian norm  $\|Z_i\|_{\psi_2}$  is bounded by a constant (depending on  $N_I/N$ ), independent of  $N$ :

$$\sup_{p \geq 1} p^{-1/2} (\mathbb{E}[Z_i^p])^{1/p} \leq \sup_{p \geq 1} p^{-1/2} (N/N_I)^2 a^2 (N_I/N)^{1/p-1} \leq (N/N_I)^2 a^2 (N_I/N)^{-1}.$$

**Proposition 2.7.** Suppose that  $\sqrt{N_I} > \sqrt{n} + t$  for  $t > 0$ . Then with high probability

$$\frac{1}{\sqrt{2}} \|x_{min} x_{min}^\top - x_0 x_0^\top\| \leq (N_I/N) c_7 \sqrt{\pi/6}.$$

for some constant  $c_7 > 0$ , independent of  $N$ .

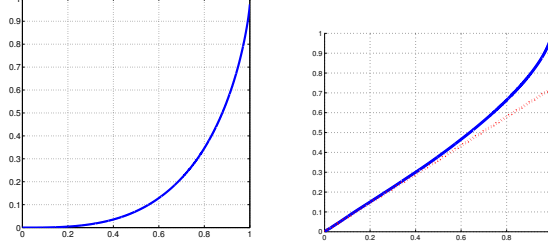


Figure 4: Figures show  $N^{-1} \|b_I\|^2$  and  $N_I^{-1/2} \|b_I\|$  vs.  $N_I/N$ . The red dotted line is  $\sqrt{\pi/6} (N_I/N)$  vs.  $N_I/N$ .

*Remark 2.8.* The following simulation illustrates that  $x_{min}$  is a good initialization  $x^0$ . Use the alternating minimization of  $x$  and  $s$  to solve the problem

$$\min_s \min_x \|Ax - sb\|^2, \quad |s| = 1.$$

Choose  $A$  to be a Gaussian random matrix from  $\mathbf{R}^{240 \times 60}$  and  $N_I = 90$ . Rescale both  $x_0, x^*$  to unit vectors, where  $x^*$  is the vector  $x$  at the final iteration. In Fig. 5, the reconstruction error is measured in terms of

$$\|x_0 x_0^\top - x^* x^{*\top}\|.$$

The figures in Fig. 5 show the results of 100 trials using two different initializations. Obviously the singular vector is a good initialization.

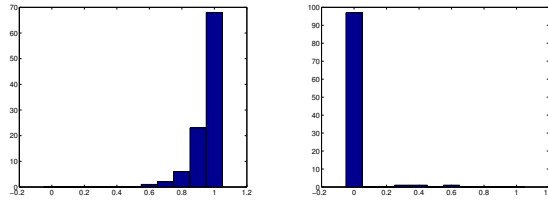


Figure 5: Figures show the histogram of the reconstruction error via the random initialization (left) and via the  $x_{min}$  initialization (right).

## 2.4 ADM with rank- $r$

Under some circumstances, the singular vector corresponding to the least singular value becomes a poor initialization for  $x_0$ , for instance, in the presence of noise. Empirically, we find that the ADM with rank  $r$  can alleviate the situation; see 3.2.

We propose the rank- $r$  method:

$$\min_{x \in \mathbf{R}^{n,r}} \frac{1}{2} \| |Ax| - b \|^2, \quad (2.10)$$

where  $|z|$  refers to the vector whose  $i$ -th entry is the vector norm of the  $i$ -th row of the matrix  $z$ , i.e.,  $(\sum_{j=1}^r z_{i,j}^2)^{1/2}$ . Note that when  $r = n$ , the set  $\{xx^\top : |Ax| = b\}$  is convex. In practical applications, we consider  $r < n$  to save the computational load. Hence, instead of vectors  $x, z$  in Eq. (2.1), we consider matrices  $x \in \mathbf{R}^{n,r}$  and  $z \in \mathbf{R}^{p,r}$  with  $Ax = z$  in the non-convex minimization problem,

$$\min_{z,x} L(z, x, \lambda), \quad (2.11)$$

$$L(z, x, \lambda) := \left\{ \frac{1}{2} \| |z| - b \|^2 + \text{tr}(\lambda^\top (Ax - z)) + \frac{\beta}{2} \| Ax - z \|_F^2 \right\}. \quad (2.12)$$

Similar to Alg. 2.1, we can adopt the ADM consisting of  $z, x, \lambda$ -iterations to solve the non-convex minimization problem. With  $\lambda$  fixed, the optimal matrices  $z, x$  have the following explicit expression.

**Proposition 2.9.** *Suppose  $(z, x)$  is a minimizer in Eq. (2.11); then,  $(zV, xV)$  is also a minimizer for any orthogonal matrix  $V \in \mathbf{R}^{n,n}$ . Moreover,*

- for each  $z$  fixed,  $x = A^\dagger(z - \beta^{-1}\lambda)$  is the optimal matrix.
- Fixing  $x$ , write  $u = Ax + \beta^{-1}\lambda$ , then the optimality of  $z$  is

$$z = \frac{u}{|u|} \frac{b + \beta|u|}{1 + \beta}.$$

*Proof.* We only prove the  $z$ -part. The  $x$ -part is obvious. Because  $L$  is separable in each row  $z_i$  of  $z$ , then the optimization of  $z_i$  can be solved via

$$\min_{z_i} \frac{1}{2} (\|z_i\|^2 - 2b_i\|z_i\| + b_i^2) + \frac{\beta}{2} (\|u_i\|^2 - 2u_i \cdot z_i + \|z_i\|^2).$$

The optimality of  $z_i$  occurs if and only if  $z_i$  parallels  $u_i/\|u_i\|$ . Let  $z_i = \alpha_i u_i/\|u_i\|$  with  $\alpha_i$  to be determined. Thus,

$$\min_{\alpha_i} \frac{1}{2} (\alpha_i^2 - 2b_i|\alpha_i| + b_i^2) + \frac{\beta}{2} (\|u_i\|^2 - 2\alpha_i\|u_i\| + \alpha_i^2).$$

Then,  $\alpha_i \geq 0$  and

$$\alpha_i - b_i + \beta(\alpha_i - \|u_i\|) = 0,$$

i.e.,  $\alpha_i = (1 + \beta)^{-1}(b_i + \beta\|u_i\|)$  completes the proof.  $\square$

Empirical experimentation shows that the above ADM can usually yield an optimal solution  $x \in \mathbf{R}^{n,r}$  with rank not equal to one, which does satisfy  $|Ax| = b$ . To recover the rank-one matrix  $x_0 x_0^\top$ , we take the following steps. First, we standardize  $A$  to be a matrix consisting of orthogonal columns via QR or SVD factorizations, such that  $I_{n \times n}$  lies in the range of  $\mathcal{A}$ . Indeed,

$$\sum_{i=1}^N a_i a_i^\top = A^\top A = I_{n \times n}.$$

When  $y \in \mathbf{R}^{n,r}$ , we have  $b^2 \cdot e = \text{tr}(yy^\top) = \|y\|_F^2$ . Hence, the norm  $\|y\|_F^2 = \sum_{i=1}^n \sigma_i(y)^2$  remains constant for all the feasible solutions  $\{y : |Qy| = b\}$ , where  $\sigma_i(y)$  refers to the singular values of  $y$ . Second, consider the objective function to retrieve the matrix  $y$  with the maximal leading singular value,

$$\min_y \left( \frac{1}{2} \|Qy - z\|_F^2 - \gamma \sigma_1(y) \right), \quad (2.13)$$

where  $\gamma > 0$  is some parameter to balance the fidelity  $|Qy| = |z| = b$  and the maximization of the leading singular value  $\sigma_1(y)$ .<sup>5</sup> Since the leading singular value of  $y$  is maximized, there is no guarantee that we can always obtain the global optimal solution.

**Proposition 2.10.** Write  $Q^\top z$  in the SVD factorization,

$$Q^\top z = U_z D_z V_z^\top.$$

Then the optimal matrix  $y$  in Eq. (2.13) is  $y = U_z D_y V_z^\top$  in the SVD factorization, where  $D_y = D_z + \gamma e_1 e_1^\top$ .

*Proof.* Observe that

$$\frac{1}{2} \|Qy - z\|_F^2 - \gamma \sigma_1(y) = \frac{1}{2} \|U_z^\top Q U_y D_y V_y^\top V_z - D_z\|_F^2 - \gamma D_y(1, 1),$$

where  $D_y(1, 1)$  refers to the (1,1) entry of the diagonal matrix  $D_y$ . Due to the rotational invariance of the Frobenius norm, then the first term achieves its minimum when

$$U_z^\top Q U_y D_y V_y^\top V_z = D_z + \alpha e_1 e_1^\top \quad (2.14)$$

and  $\alpha$  is the minimizer of

$$\min_{\alpha} \left( \alpha^2/2 - \gamma(D_z(1, 1) + \alpha) \right), \text{ i.e., } \alpha = \gamma.$$

Also, Eq. (2.14) yields  $U_z^\top Q U_y = I$  and  $V_y^\top V_z = I$ , which completes the proof.  $\square$

*Remark 2.11.* Suppose that  $|Ax| = b$  for some  $x \in \mathbf{R}^{n,r}$ . Then, the minimizer of

$$\min_x \left( \frac{1}{2} \| |Ax| - b \|^2 - \gamma \sigma_1(x) \right)$$

---

<sup>5</sup> In experiments, we choose  $\gamma = 0.01$ .

is  $(1 + \gamma)x_0$ . Indeed, consider  $x = \alpha x_0$ . Then,

$$\alpha = \arg \min_{\alpha} \left( \frac{1}{2}(\alpha - 1)^2 \|b\|^2 - \gamma \alpha \right) = 1 + \gamma.$$

In Eq. (2.11), replacing  $A$  with  $Q$  and replacing the term  $\frac{\beta}{2} \|Ax - z\|_F^2$  with

$$\beta \left( \frac{1}{2} \|Qy - z\|_F^2 - \gamma \sigma_1(y) \right),$$

then we adopt the ADM to retrieve a rank-one solution.

**Algorithm 2.12.** *Initialize a random matrix  $y \in \mathbf{R}^{n,r}$  and  $\lambda^0 = 0_{N,r} \in \mathbf{R}^{N,r}$ . Repeat the following steps,  $k = 1, 2, \dots$ . Then let the solution  $x^*$  be the first column of  $U_z$ , i.e., the singular vector corresponding to the maximal singular value.*

1. *z-iteration:*

$$u = Qy^k + \lambda^k \beta^{-1}, \quad z^{k+1} = \frac{u}{|u|} \frac{b + \beta|u|}{1 + \beta},$$

2.  *$\lambda$ -iteration:*

$$\lambda^{k+1} = \lambda^k + \beta(Qy^k - z^{k+1}),$$

3. *y-iteration:*

$$U_z D_z V_z = Q^\top (z^{k+1} - \lambda^{k+1} \beta^{-1}), \quad y^{k+1} = U_z (D_z + \gamma e_1).$$

## 2.5 Standardized frames with equal norm

In the simulations (section 3.1), we will show the importance of the unit norm condition  $\|a_i\| = 1$  for  $i = 1, \dots, N$  in the ADM approach. When the QR factorization is used to generate an equivalent standardized matrix consisting of rows  $\{a_i\}_{i=1}^N$ , the sensing vectors  $\{a_i\}$  do not have equal norm in general.

The following theorem states that we can standardize  $A$  to obtain an orthogonal matrix  $Q$  whose rows have equal norm. The proof is given in the appendix.

**Theorem 2.13.** *Given a matrix  $A \in \mathbf{R}^{N \times n}$  satisfying the rank\* condition and  $N > n$ , we can find a unique diagonal matrix  $D$  with  $D_{i,i} > 0$ , such that*

$$D^{-1/2} A = QB,$$

and  $Q$  is one standardized matrix, which is one projection matrix with  $Q^\top Q = I_{n \times n}$ ,  $(QQ^\top)_{i,i} = (n/N)$  for all  $i$ , where  $B$  is some  $n \times n$  non-singular matrix.

Here the diagonal value  $n/N$  is the average of the norm  $\|Q\|_F^2 = \text{tr}(Q^\top Q) = \text{tr}(QQ^\top) = n$ . Also,

$$N = \|D^{1/2} A\|_F^2 = \|QB\|_F^2 = \|B\|_F^2.$$

With the uniqueness of  $D$ ,  $Q$  is also determined uniquely up to the right multiplication of an orthogonal matrix. Indeed,  $B$  is uniquely determined up to the left multiplication of an orthogonal matrix:

$$A^\top D^{-1/2} D^{-1/2} A = B^\top Q^\top QB = B^\top B.$$

Recall that  $A$  satisfies the rank\* condition if any square  $n$ -by- $n$  sub-matrix of  $A$  is full rank. When a matrix  $A$  satisfies the rank\* condition then there exists no orthogonal matrix  $V \in \mathbf{R}^{n \times n}$ , such that

$$AV = C = \begin{pmatrix} C_{1,1} & 0 \\ 0 & C_{2,2} \end{pmatrix}, \quad (2.15)$$

where the 0s refer to zero sub-matrices with size  $(N - N_1) \times n_1$  and size  $N_1 \times (n - n_1)$  and  $C_{1,1}$  is an  $N_1 \times n_1$  matrix.<sup>6</sup> Furthermore, the condition ensures that the norm of each row must be positive. It is easy to see that, with probability one, Gaussian random matrices satisfy the rank\* condition.

### 3 Experiments

#### 3.1 ADM failure experiments

Due to the nature of nonconvex minimization, the algorithm can fail to converge, which is indeed observed in the following two simulations.

First, let us denote the input data by  $(A, b)$  with  $b_i \neq 0$  and the unknown signal by  $x_0$ . Mathematically, solving problem (i)

$$|Ax_0| = b$$

is equivalent to solving problem (ii)

$$b_i^{-1}|a_i \cdot x_0| = 1.$$

However, solving these two problems via the ADM [22] can yield different results.

Let  $A$  be a real Gaussian random matrix,  $A \in \mathbf{R}^{N \times n}$ . Let  $b = |Ax_0|$ . Rescale the system by  $b^{-1}$ , i.e., the input data becomes  $(b^{-1}A, 1_{N \times 1})$ , thus equal measurement values. Figure 6 shows the error  $\| |AA^\dagger z| - b \|$  at each iteration. Here we use the random initialization for  $x^0$ .

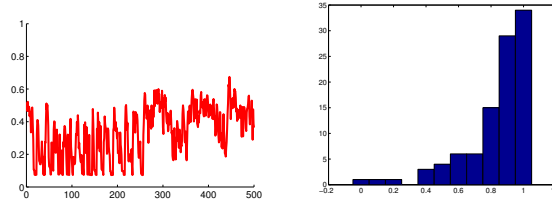


Figure 6: The left figure shows the error  $\| |AA^\dagger z| - b \|$  vs. the number of iteration via ADM with rank one [22]. The right figure shows the histogram  $\|x_0 x_0^\top - x^* x^{*\top}\|$  of 100 reruns.

<sup>6</sup> Otherwise, it is easy to see that one of the following submatrices must be rank deficient: (1) the top submatrix with entries  $\{C_{i,j} : i, j = 1, \dots, n\}$  or (2) the bottom submatrix with entries  $\{C_{i,j} : i = N - n + 1, \dots, N, j = 1, \dots, n\}$ .



Second, we demonstrate a few experiments where the ADM also fails to converge. The convergence failure sheds light on the importance of the two proposed assumptions in Section 2.2.

We sort a set of random generated sensing vectors  $\{a_i \in \mathbf{R}^{100}\}_{i=1}^{400}$ , such that

$$|a_i \cdot x_0| \leq |a_j \cdot x_0| \text{ for all } i < j.$$

That is, the indices are sorted according to the values  $b_i$ . We consider three different manners of selecting 200 sensing vectors  $\{a_i\}$ : (1) the vectors with the smallest indices, (2) the vectors with the largest indices, and (3) a combination with 199 small indices and one large index. Finally, we compare these results with the result using a random selection of sensing vectors, as shown in Fig. 7. Here, we fix rank  $r = 1$  and  $\beta = 0.01$ . Clearly, the combination with smaller indices and larger indices performs best.

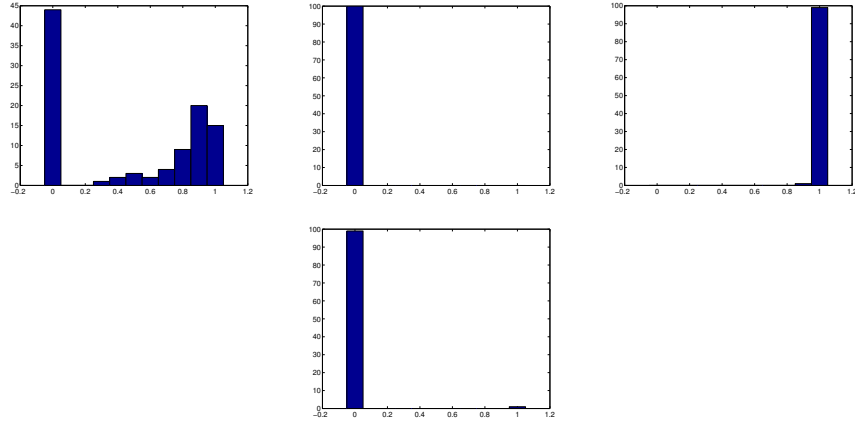


Figure 7: Figure shows the histogram  $\|x_0 x_0^\top - x^* x^{*\top}\|$  under four different sets of  $\{a_i\}_{i=1}^{400}$ . In the top row, Left, middle and right subfigures show the results with  $\{a_i\}_{i=1}^{200}$ , with  $\{a_i\}_{i=1}^{199} \cup \{a_{400}\}$  and with  $\{a_i\}_{i=201}^{400}$ . The bottom subfigure shows the result when randomly sampling 200 sensing vectors.

### 3.2 Comparison experiments with noises

In this subsection, we demonstrate the performance of the ADM with  $r = 1$  and  $r > 1$  on a number of simulations, where Gaussian white noise is added. The noise-corrupted data,  $b$ , is generated,

$$b^2 = \max((Ax_0)^2 + \text{noise}, 0).$$

The signal-to-noise ratio is defined by

$$SNR = 10 \log_{10} \frac{\|Ax_0\|^2}{\|noise\|_F}.$$

In Fig. 8, we consider  $A$  to be a real Gaussian random matrix with  $N = 2n$ . We rerun the experiments 200 times to test the effect of random initialization. The first row shows the histogram result with  $n = 30$  and  $\text{noise} = 0$ . All the algorithms with  $r = 1, 2$ , and 3 work well. The second row shows the histogram result with  $n = 30$  and  $\text{SNR} = 29$ . Here, we use  $\beta = 0.001$ . Obviously the algorithms with  $r > 1$  have better performances.

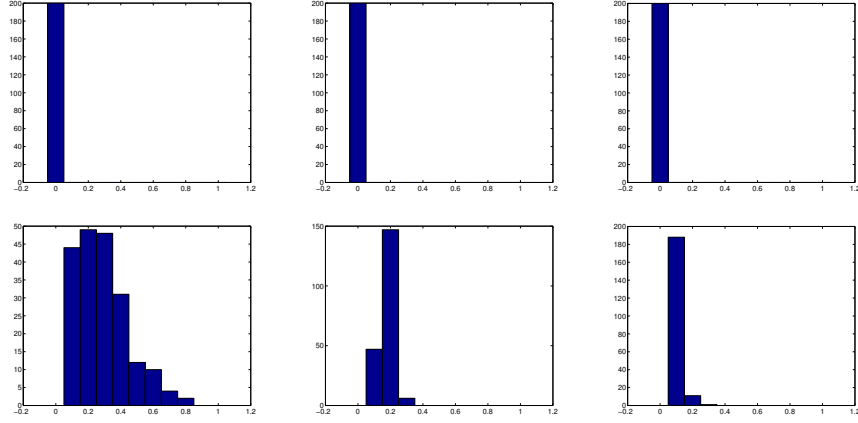


Figure 8: Figure shows the histogram  $\|x_0 x_0^\top - x^* x^{*\top}\|$  under the noise effect. Left, middle and right columns show the results with rank  $r = 1$ ,  $r = 2$  and  $r = 3$ . Here we use the random initialization.

Let  $n = 30$ ,  $N = 3n$  with  $\beta = 0.01$ . In Fig. 9, we demonstrate the comparison between the random initialization and the singular vector initialization, i.e., the initialization is chosen to be the singular vector corresponding to the least singular value of  $A_I \in \mathbf{R}^{45 \times 30}$ . Data  $b$  is generated with  $\text{noise} = 2 \times 10^{-4} \times \text{Normal}(0, 1)$ ,  $\text{SNR} = 25\text{dB}$ . Furthermore, with the presence of noise, when ADM with  $r = 1$  is employed, the difference between the two initializations is very little, in contrast to the simulation result shown in Remark 2.8.

### 3.3 Phase retrieval experiments

Next, we report phase retrieval simulation results (Fourier matrices), with  $x_0$  being real, positive images. Images are reconstructed subject to the positivity constraints (i.e., the leading singular vector). The results are provided to show some advantage of ADM with  $r = 2$  over ADM with  $r = 1$ . Here we use  $\beta = 0.1$  in the following experiment.

According to our experience, the phase retrieval with the Fourier matrix is a very difficult problem, in particular in the presence of noise. To alleviate the difficulty, researchers have suggested random illumination to enforce the uniqueness of solutions [9]. It is known that the phase retrieval has a unique solution up to three classes: constant global phase, spatial

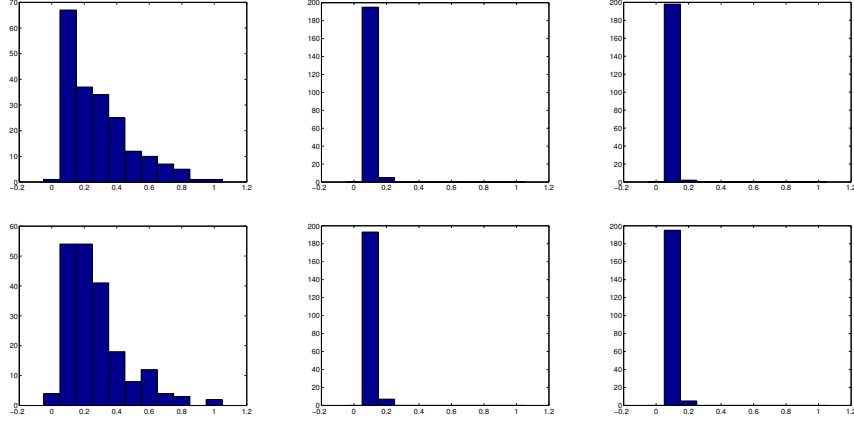


Figure 9: Figure shows the histogram  $\|x_0 x_0^\top - x^* x^{*\top}\|$  of 200 trials under the noise effect. Left, middle and right columns show the results with rank  $r = 1$ ,  $r = 2$  and  $r = 3$ . Here, we use the proposed initialization in the top row and the random initialization in the second row.

shift, and conjugate inversion. With high probability absolute uniqueness holds with a random phase illumination; see Cor. 1 [9]. Our experiences show that the random phase illumination works much better than the above uniform illumination.

In Fig. 10, we demonstrate the the ADM with  $r = 1, 2$  on the images with random phase illumination. Let  $x_0 \in \mathbf{R}^{300 \times 300}$  be the intensity of the Lena image<sup>7</sup>, see the bottom subfigure. We add noise and generate the data

$$b^2 = \max(|Ax_0|^2 + \text{noise}, 0),$$

where  $A$  is the Fourier matrix. The SNR is  $39.8dB$  and the oversampling is 1.23. Reconstruction errors  $\|\frac{x^*}{\|x^*\|_F} - \frac{x_0}{\|x_0\|_F}\|_F$  for rank one and rank two are 0.126 and 0.109, respectively. The ADM with  $r = 2$  has a better reconstruction.

### 3.4 Conclusions

In this paper, we discuss the rank-one matrix recovery via two approaches. First, the rank-one matrix is computed among the Hermitian matrices as in PhaseLift. We make the observation that matrices in the feasible set have equal trace norm via the measurement matrices with orthonormal columns. Experiments show that with the aid of these orthogonal frames, exact recovery occurs under a smaller  $N/n$  ratio compared with the PhaseLift in both real and complex cases. In the second part of the

<sup>7</sup>We downsample the Lena image from <http://www.ece.rice.edu/~wakin/images/> by approximately a factor 2 and use zero padding with the oversampling rate [17][15] 1.23.

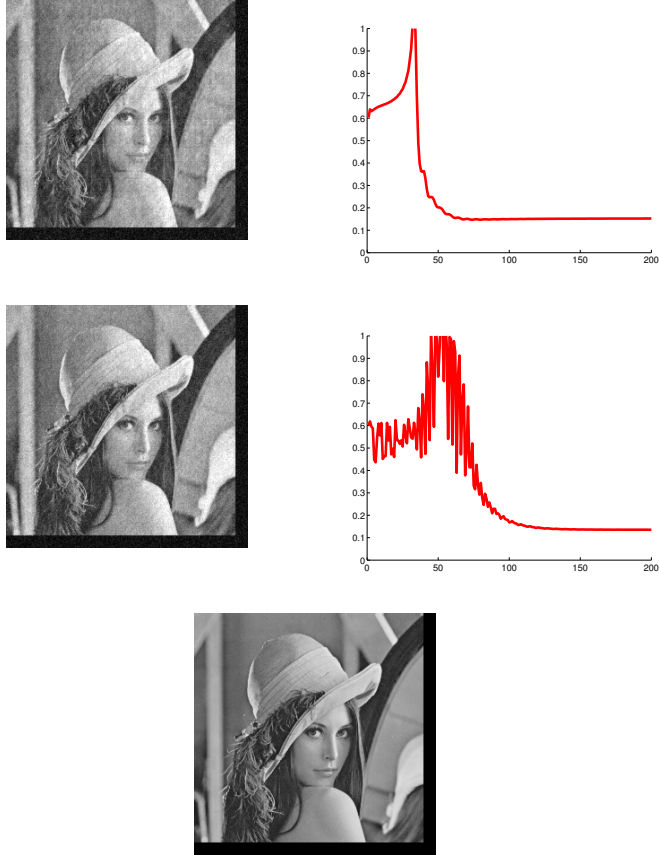


Figure 10: Figures show the reconstructed images and the error  $\|AA^\dagger z - b\|_F$  vs. the number of iteration via ADM with rank  $r = 1$  (the first row) and  $r = 2$  (the second row).

paper, we discuss the “lifting” of the nonconvex alternating direction minimization method from rank-one to rank- $r$  matrices,  $r > 1$ . The benefit of this relaxation cannot be overestimated, because the construction of large Hermitian matrices is avoided, as is the associated Hermitian matrices projection. Comparing with the ADM with rank-one, the ADM with rank  $r > 1$  performs better in recovering noise-contaminated signals, which is demonstrated in simulation experiments.

Another contribution is the error estimate between the unknown signal and the singular vector corresponding to the least singular value. The initialization has an effect of importance in the nonconvex minimization. We demonstrate that a good initialization can be the least singular vector of the subset of sensing vectors corresponding to the small measurement values  $b_i$ . In the case of real Gaussian matrices, the error can be reduced, as the number of measurements grows at a rate proportional to the dimension of unknown signals. One of our future works is the generalization of the error estimate to complex frames, in particular the case of the Fourier matrix.

## A Standardization of $A$

In the following, we will prove Theorem 2.13 in several steps. We discuss the existence first. The uniqueness analysis will be shown later. Fixing  $A$ , let  $\mathcal{D}$  be the inverse matrices of diagonal matrices  $D$ ,

$$\mathcal{D} := \{D^{-1} \in \mathbf{R}^{N \times N} : \|D^{-1/2}A\|_F^2 = \sum_{i=1}^N D_{i,i}^{-1} \sum_{j=1}^n A_{i,j}^2 = N, D_{i,i} \geq 0\}.$$

Clearly  $\mathcal{D}$  is nonempty and convex compact. In fact,  $D_{i,i}$  has a positive lower bound,

$$D_{i,i} \geq N^{-1} \sum_{j=1}^n A_{i,j}^2 \text{ for all } i.$$

For each  $D^{-1} \in \mathcal{D}$ , let  $f : \mathcal{D} \rightarrow \mathcal{D}$  be the function

$$f(D^{-1}) = \hat{D}^{-1}, \text{ where } QB = D^{-1/2}A \text{ is the QR factorization,}$$

and each row of  $\hat{D}^{-1/2}AB^{-1}$  has norm one. In fact, the function  $f$  generates iterations  $\{(D^k)^{-1}\}_{k=0}^\infty$  with  $(D^{k+1})^{-1} = f((D^k)^{-1})$ . That is, start with  $Q^0 = A$ . Repeat the two steps for  $k = 0, 1, 2, \dots$  until it converges:

- (ii) Normalize the row of  $Q^k$  by  $(D^k)^{-1/2}Q^k$ ;
- (ii) Take the QR factorization:  $(D^k)^{-1/2}Q^{k-1} = Q^k R^k$ .

Since  $D^{-1/2}A$  has rank  $n$ , then  $B$  has rank  $n$  and  $B^{-1}$  exists. The function  $f$  is well defined: Once  $B$  is given, then choose the diagonal matrix  $D$  to be that which normalizes the rows of  $AB^{-1}$ . According to Brouwer’s fixed-point theorem, we have the existence of  $D$ , such that  $D^{-1/2}AB^{-1} = Q$  consists of orthogonal columns and each row has norm one.

Before the uniqueness proof, we state one equation of  $D$ .

**Proposition A.1.** *The diagonal matrix  $D$  satisfies the equation,*

$$(n/N)D_{i,i} = (A(A^\top D^{-1}A)^{-1}A^\top)_{i,i}. \quad (\text{A.1})$$

*Proof.* According to  $D^{-1/2}A = QB$ , we have

$$(n/N)D_{i,i} = (D^{1/2}QQ^\top D^{1/2})_{i,i} = (A(B^\top B)^{-1}A^\top)_{i,i}.$$

Note that  $B^\top B = A^\top D^{-1/2}Q^\top QD^{-1/2}A = A^\top D^{-1}A$ . Thus,

$$(n/N)D_{i,i} = (A(A^\top D^{-1}A)^{-1}A^\top)_{i,i}.$$

□

**Proposition A.2.** *Let  $p_i \in (0, 1)$ ,  $i = 1, \dots, n$  with  $\sum_{i=1}^n p_i = 1$ . Let  $\lambda_i > 0$  for  $i = 1, \dots, n$ . Then*

$$\sum_{i=1}^n p_i \lambda_i \geq \left( \sum_{i=1}^n p_i \lambda_i^{-1} \right)^{-1},$$

where equality holds if and only if  $\{p_i\}_{i=1}^n$  are equal.

*Proof.* Let  $f(x) = x^{-1}$  for  $x > 0$ , which is strictly convex. The statement is the application of Jensen inequality,

$$\sum_{i=1}^n p_i \lambda_i^{-1} \geq \left( \sum_{i=1}^n p_i \lambda_i \right)^{-1}.$$

□

**Proposition A.3.** *Suppose that  $Q$  is a standardized matrix satisfying the  $\text{rank}^*$  condition. Let  $F^1$  be a positive diagonal matrix. Then the iteration*

$$F_{i,i}^{k+1} = (N/n)(Q(Q^\top (F^k)^{-1}Q)^{-1}Q^\top)_{i,i}, \quad k = 1, \dots$$

yields  $\lim_{k \rightarrow \infty} F^k = cI_{N \times N}$ , where  $c$  is some scalar.

*Proof.* We will show  $\text{tr}(F^{k+1}) \leq \text{tr}(F^k)$ . Suppose that  $\{(\lambda_i^{-1}, q_i)\}_i$  are eigenvalues-eigenvectors of  $Q^\top (F^k)^{-1}Q$ , then  $\{(\lambda_i, q_i)\}_i$  are eigenvalues-eigenvectors of  $(Q^\top (F^k)^{-1}Q)^{-1}$ . Hence,

$$\lambda_i^{-1} = q_i^\top Q^\top (F^k)^{-1} Q q_i,$$

and

$$\text{tr}(F^{k+1}) = \sum_{j=1}^N \mu_j^{k+1} = (N/n) \sum_{i=1}^n \lambda_i = (N/n) \sum_{i=1}^n \left( \sum_{j=1}^N (F_{j,j}^k)^{-1} (Q q_i)_j^2 \right)^{-1}.$$

Denote the  $j$ -th entry of  $|(Q q_i)_j|$  by  $p_{j,i}$ . Then  $\sum_{j=1}^N p_{j,i}^2 = 1$  and  $\sum_{i=1}^n p_{j,i}^2 = n/N$ . Let  $\{\mu_i^k\}_{i=1}^N$  be the diagonal entries of  $F^k$ . Then

$$\sum_{j=1}^N \mu_j^{k+1} = (N/n) \sum_{i=1}^n \left( \sum_{j=1}^N (\mu_j^k)^{-1} p_{j,i}^2 \right)^{-1} \leq (N/n) \sum_{i=1}^n \sum_{j=1}^N \mu_j^k p_{j,i}^2 = \sum_{j=1}^N \mu_j^k,$$

where the last equality is due to  $\sum_{i=1}^n p_{j,i}^2 = n/N$ . Hence,  $\text{tr}(F^{k+1}) \leq \text{tr}(F^k)$ . Denote one of limiting points of  $\mu_i^k$  by  $\mu_i^*$  and then

$$\left(\sum_{j=1}^N p_{j,i}^2 (\mu_j^*)^{-1}\right)^{-1} = \sum_{j=1}^N p_{j,i}^2 \mu_j^* \text{ for all } i.$$

Hence,  $\mu_i^* = \mu_j^*$  for all  $i, j$  with  $p_{j,i} > 0$ . Due to the rank\* condition,  $QV$  cannot be written in the form of Eq. (2.15) for any orthogonal matrix  $V$  whose columns are orthonormal vectors  $\{q_i\}_{i=1}^n$  with  $p_{j,i} = |(QV)_{j,i}|$ . Hence,  $c = \mu_i^* = \mu_j^*$  for all  $i, j$ .  $\square$

Finally, we complete the proof in the following.

**Proposition A.4.** *Suppose that  $A$  satisfies the rank\* condition. Let  $D_*$  be one solution of Eq. (A.1). Then with any positive diagonal matrix  $D^0$ , the iteration*

$$D^{k+1} = (N/n) \text{diag}(A(A^\top (D^k)^{-1} A)^{-1} A^\top), \quad k = 1, \dots,$$

*yields*

$$\lim_{k \rightarrow \infty} D^k = D_*.$$

*Thus,  $D_*$  is unique.*

*Proof.* Let  $D_*^{-1/2} A = QB$  be the QR factorization of  $D_*^{-1/2} A$ . Then

$$(A^\top D^{-1} A)^{-1} = B^{-1} (Q^\top (D_*^{-1/2} D D_*^{-1/2})^{-1} Q)^{-1} B^{-1},$$

and the iteration becomes

$$D_*^{-1/2} D^{k+1} D_*^{-1/2} = (N/n) \text{diag}(Q(Q^\top (D_*^{-1/2} D^k D_*^{-1/2})^{-1} Q)^{-1} Q^\top), \quad k = 1, \dots.$$

Let  $F^k = D_*^{-1/2} D^k D_*^{-1/2}$ . Since  $A$  satisfies the rank\* condition, then for any nonsingular matrix  $B$ ,  $D_*^{-1/2} A B^{-1}$  also satisfies the rank\* condition and cannot be written in the form in Eq. (2.15) for any orthogonal matrix. According to Prop. A.2, the proof is completed.  $\square$

## References

- [1] Radu Balan, Pete Casazza, and Dan Edidin. On signal reconstruction without phase. *Applied and Computational Harmonic Analysis*, 20(3):345–356, 2006.
- [2] Afonso S. Bandeira, Jameson Cahill, Dustin G. Mixon, and Aaron A. Nelson. Saving phase: Injectivity and stability for phase retrieval. *Applied and Computational Harmonic Analysis*, 2013.
- [3] Heinz H. Bauschke, Patrick L. Combettes, and D. Russell Luke. Phase retrieval, error reduction algorithm, and Fienup variants: a view from convex optimization. *J. Opt. Soc. Amer. A*, 19:1334–1345, 2002.

- [4] E. Candès, Y. Eldar, T. Strohmer, and V. Voroninski. Phase retrieval via matrix completion. *SIAM Journal on Imaging Sciences*, 6(1):199–225, 2013.
- [5] Emmanuel J. Candès and Xiaodong Li. Solving quadratic equations via phaselift when there are about as many equations as unknowns. *Foundations of Computational Mathematics*, pages 1–10, 2013.
- [6] Emmanuel J. Candès, Thomas Strohmer, and Vladislav Voroninski. PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.
- [7] Hung Chen. *Lecture note*. <http://www.math.ntu.edu.tw/~hchen/teaching/LargeSample/notes/noteorder.pdf>.
- [8] Laurent Demanet and Paul Hand. Stable optimizationless recovery from phaseless linear measurements. *Journal of Fourier Analysis and Applications*, pages 1–23, 2012.
- [9] Albert Fannjiang. Absolute uniqueness of phase retrieval with random illumination. *Inverse Problems*, 28(7):Article ID 075008, 20 p., 2012.
- [10] Albert Fannjiang and Wenjing Liao. Phase retrieval with random phase illumination. *J. Opt. Soc. Am. A*, 29(9):1847–1859, Sep 2012.
- [11] J. R. Fienup. Phase retrieval algorithms: a comparison. *Appl. Opt.*, 21(15):2758–2769, Aug 1982.
- [12] R. W. Gerchberg and W. Owen Saxton. A practical algorithm for the determination of the phase from image and diffraction plane pictures. *Optik*, 35:237–246, 1972.
- [13] Guoyong Gu, Bingsheng He, and Xiaoming Yuan. Customized proximal point algorithms for linearly constrained convex minimization and saddle-point problems: a unified approach. *Computational Optimization and Applications*, pages 1–27, 2013.
- [14] N.E. Hurt. *Phase Retrieval and Zero Crossings: Mathematical Methods in Image Reconstruction*. Mathematics and Its Applications. Springer, 2001.
- [15] J. Miao, D. Sayre, and H. N. Chapman. Phase retrieval from the magnitude of the fourier transforms of nonperiodic objects. *J. Opt. Soc. Am. A*, 15(6):1662–1669, Jun 1998.
- [16] Jianwei Miao, Tetsuya Ishikawa, Qun Shen, and Thomas Earnest. Extending x-ray crystallography to allow the imaging of noncrystalline materials, cells, and single protein complexes. *Annual Review of Physical Chemistry*, 59(1):387–410, 2008. PMID: 18031219.
- [17] R. P. Millane. Phase retrieval in crystallography and optics. *J. Opt. Soc. Am. A*, 7(3):394–411, Mar 1990.
- [18] Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices: extreme singular values. *Proceedings of the International Congress of Mathematicians*, 2010.



- [19] Yoav Shechtman, Yonina C. Eldar, Oren Cohen, Henry N. Chapman, Jianwei Miao, and Mordechai Segev. Phase retrieval with application to optical imaging. *CoRR*, abs/1402.7350, 2014.
- [20] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *CoRR*, abs/1011.3027, 2010.
- [21] G.A. Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra and its Applications*, 170:33 – 45, 1992.
- [22] Zaiwen Wen, Chao Yang, Xin Liu, and Stefano Marchesini. Alternating direction methods for classical and ptychographic phase retrieval. *Inverse Problems*, 28(11):115010, 2012.